Managing the Privacy-loss Budget for the 2020 Census

John M. Abowd and Victoria Velkoff
Associate Director R&M, Chief Scientist and Associate Director Demographic Programs
U.S. Census Bureau

Census Scientific Advisory Committee
March 28, 2019

Update on Reconstruction and Re-identification

- As presented to the American Association for the Advancement of Science (AAAS) on February 16, 2019
- Technical paper under internal review prior to submission for external peer review

What We Did

- Database reconstruction for all 308,745,538 people in 2010 Census
- Link reconstructed records to commercial databases: acquire PII
- Successful linkage to commercial data: putative re-identification
- Compare putative re-identifications to confidential data
- Successful linkage to confidential data: confirmed re-identification
- Harm: attacker can learn self-response race and ethnicity

What We Found

- Census block and voting-age correctly reconstructed in all 6,207,027 inhabited blocks
- Block, sex, age, race, ethnicity reconstructed
  - Exactly: 46% of population (142 million of 308,745,538 records in CEF)
  - Allowing age +/- one year: 71% of population (219 million of 308,745,538)
- Block, sex, age linked to commercial data to acquire PII
  - Putative re-identifications: 45% of population (138 million of 308,745,538)
- Name, block, sex, age, race, ethnicity compared to confidential CEF
  - Confirmed re-identifications: 38% of putative (52 million; 17% of population)
- For the confirmed re-identifications, race and ethnicity are learned exactly, not statistically

Schedule of Publications

- Reapportionment (December 31, 2020) *unaffected by differential privacy*
- Redistricting (PL94-171, March 31, 2021)
- Citizen Voting-Age Population (CVAP, March 31, 2021) *CVAP is not an official 2020 product*
- Standard Data Products (Spring 2021-Summer 2023)
  - Using OMB standard race and ethnicity groups
  - Including complex join queries (household-person tables)
  - Using detailed race and ethnicity categories
  - Using detailed American Indian and Alaska Native categories
- Public-use microdata (after all other releases)

Many Historical Invariants

- An invariant is published as-enumerated (no confidentiality protection)
- There is only one Constitutional invariant: reapportionment
- There are no statutory invariants
  - Confidentiality protection applies to all products
- Historically there were many invariants (2010 examples below):
  - Total population at all geographic levels
  - Voting-age population at all geographic levels
  - Number of housing units at all geographic levels
  - Number of occupied housing units at all geographic levels
  - Number and type of group quarters at all geographic levels
    - Detail in type of group quarters varies by geographic level

2018 E2E Test and 2020 Census Invariants

- Invariants in the 2018 End-to-End Census Test:
  - Total population of Providence, RI (only county tested)
  - Number of housing units at all geographic levels
  - Number of occupied housing units at all geographic levels
  - Number and type of group quarters at all geographic levels
    - Table P-42 had only 7 group quarters types
- DSEP sets the final invariants

Invariants Massively Complicate the Problem

- Internal research shows
  - Population invariants at the block and tract level were major contributors to the accuracy of the reconstruction-abetted re-identification experiments run on the 2010 Census
  - Protecting confidentiality and maintaining fitness-for-use require removing invariants at the block and tract levels
- Every invariant results in a compromise of the confidentiality protections: some plausible attack strategies are advantaged more than the formal privacy-loss parameter allows
- Formal privacy guarantees are strongest when there are no invariants and the privacy-loss parameter is used to control accuracy (see Dan Kifer talk distributed with CSAC materials)

Managing a Global Privacy-loss Budget

- There are three generic uses of the global privacy-loss budget
  - Person-level queries
    - Bulk of PL94-171 and Citizen Voting-Age Population (CVAP) tables
    - Many Demographic and Housing Characteristics (DHC) tables
    - Some tables using detailed race and ethnicity, AIAN
  - Household-level queries
    - One PL94-171 table, no CVAP tables
    - Many DHC tables
    - Most tables in detailed race, ethnicity and AIAN products
  - Household-person queries
    - None in PL94-171 nor CVAP
    - Balance of tables in DHC
- Public-use microdata would be developed from these queries, so there is no additional privacy-loss

Allocating Privacy Loss across Sets of Tables

- Requires treating the entire confidential database (CEF) as relational with hierarchy-defined relations (see Michael Hay talk, distributed with CSAC materials)
- Requires implementing privacy-loss accounting for the entire database not just separate components like person tables (PL94-171)
- Current policy: person is primary (the privacy-loss budget provides guarantees to each person in the United States)
- Privacy-loss accounting manages the budget over persons, household and household-person joins

Allocating Privacy Loss to Household and Person Tables

- Mostly solved problems
  - PL94-171, CVAP
  - Can be combined with person-level tables in DHC
  - Basic analysis was presented at the December 6, 2018 CSAC meeting
- Tractable problems
  - Balance of person tables in DHC
  - Household tables in DHC
- Remaining problems
  - Optimizing the allocation of privacy loss across the geographic hierarchy
  - Implementing improved strategies for other variables (age, OMB race)
  - Optimizing overall workload

Allocating Privacy Loss to Household-Person and Sparse Tables

- Household-person join queries are challenging
  - Computation of the sensitivity must be correctly automated
  - Privacy-loss accounting must be properly implemented
  - Resulting protected tables cannot be accurately represented with microdata
  - Requires computing published tables from protected summaries instead
- Sparse queries are also challenging
  - Detailed race, ethnicity and AIAN tables historically applied to very small populations in select geographies
  - Requires data-dependent algorithms that are not yet implemented or tested
  - Even with these algorithms, the volume of data previously published has set very difficult expectations

The Importance of Formal Privacy

- Block-level summary data from the decennial census have a long history, an important and valid use case, and can be delivered with the current formal privacy system, as demonstrated in the 2018 End-to-End Census test
- Abandoning formal privacy for the balance of 2020 Census publications exposes the entire set of publications, including the block-level tables, to the same reconstruction-abetted re-identification attack strategy to which the 2010 Census was vulnerable
- The current environment is equivalent to exposing a major cybersecurity vulnerability: you can't patch one part and leave other parts exposed—you have to fix the whole system

Questions for CSAC

- How should the Census Bureau communicate the vulnerabilities that invariants produce while trying to eliminate them from the publications?
- How can the Census Bureau effectively communicate to users that complete accuracy of inputs to their use cases is infeasible, and was not true historically?
- How can the Census Bureau best do principled balancing of the accuracy requirements of diverse use cases?
- In tuning the full geographic hierarchy, which levels make the most sense to optimize for accuracy?
- If the only feasible algorithms for producing household-person join tables and detailed race, ethnicity and AIAN tables cannot deliver microdata for tabular publication, should the Census Bureau invest in a dissemination system that publishes from protected tables instead?
- How should the Census Bureau assess the use case for PUMS and restricted-access to the confidential microdata?
- Should the Census Bureau relax the requirement that all published tables be fully consistent, as other national statistical offices have done for their census publication?
- How can the Census Bureau incorporate systems that will give a holistic perspective on the impact of these changes?

Thank you.

John.Maron.Abowd@census.gov and Victoria.A.Velkoff@census.gov